Ceph 存储系统中常见问题和排查手段

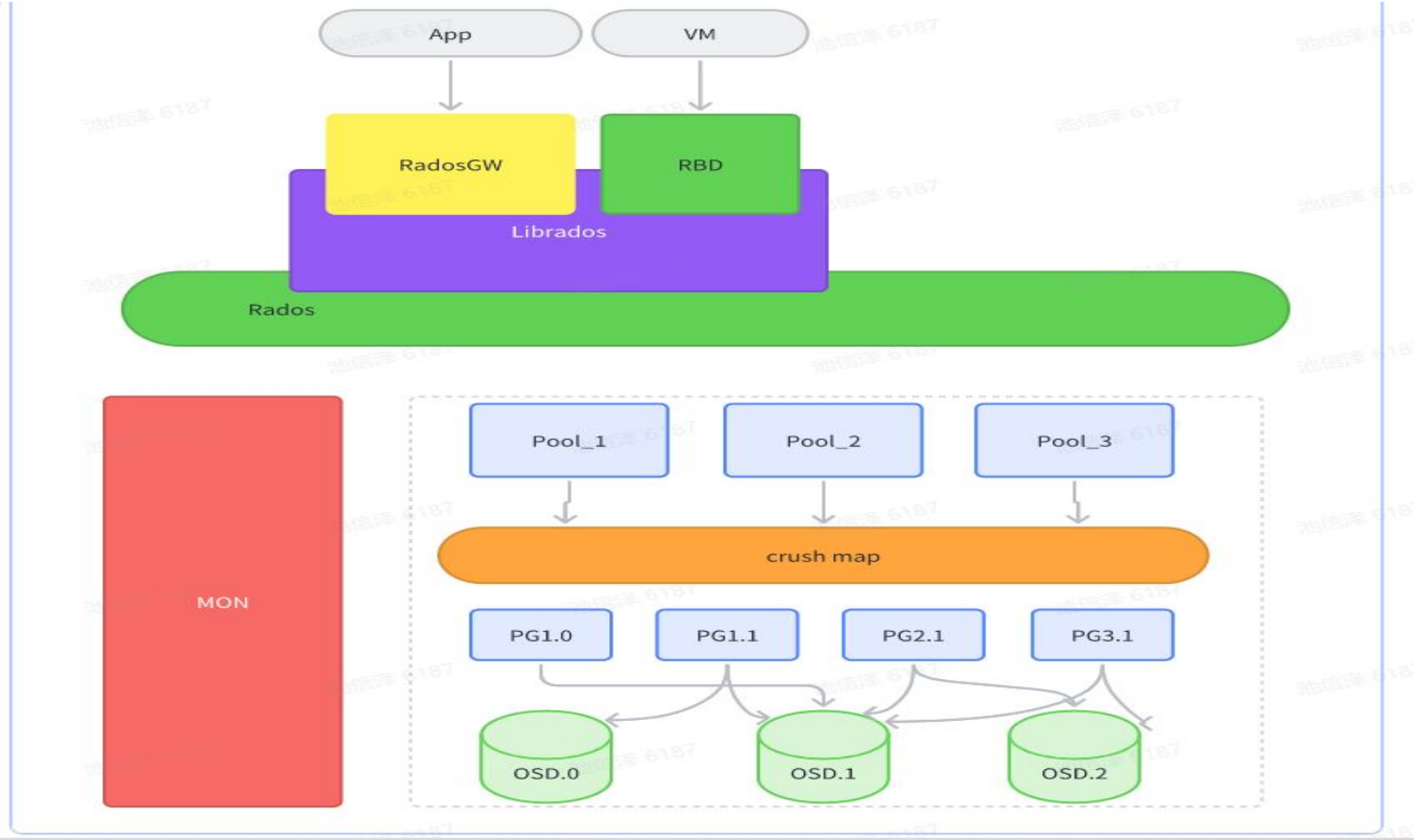
星辰天合数据科技有限公司

池信泽

Ceph 存储系统中常见问题和排查手段

- 扩缩容场景
- 集群异常
- 缓存因素
- 业务场景

Ceph 架构介绍


App
VM
RadosGW
RBD
Librados
Rados
MON
Pool_1
Pool_2
Pool_3
crush map
PG1.0
PG1.1
PG2.1
PG3.1
OSD.0
OSD.1
OSD.2

扩缩容场景

常见的问题

- PG 数量的影响
  - 容量不均衡
  - 性能不均衡
- 数据可靠性
  - 直接缩容节点数据可靠性受影响

解决方案

- 及时做PG 分裂
- 调整OSD crush 权重，预估调整后OSD 容量是否均衡
- 利用upmap 干预crush 算法
  - 手动迁移 PG
- 先预先迁移数据再缩容

集群异常

常见的问题

- 恢复卡前端IO
  - 主没有全量数据
- 单对象数据量太大
  - 桶中对象数量
- 快照太多
  - 元数据访问低下
- 多磁盘扇区损坏
  - 集群不可用

解决方案

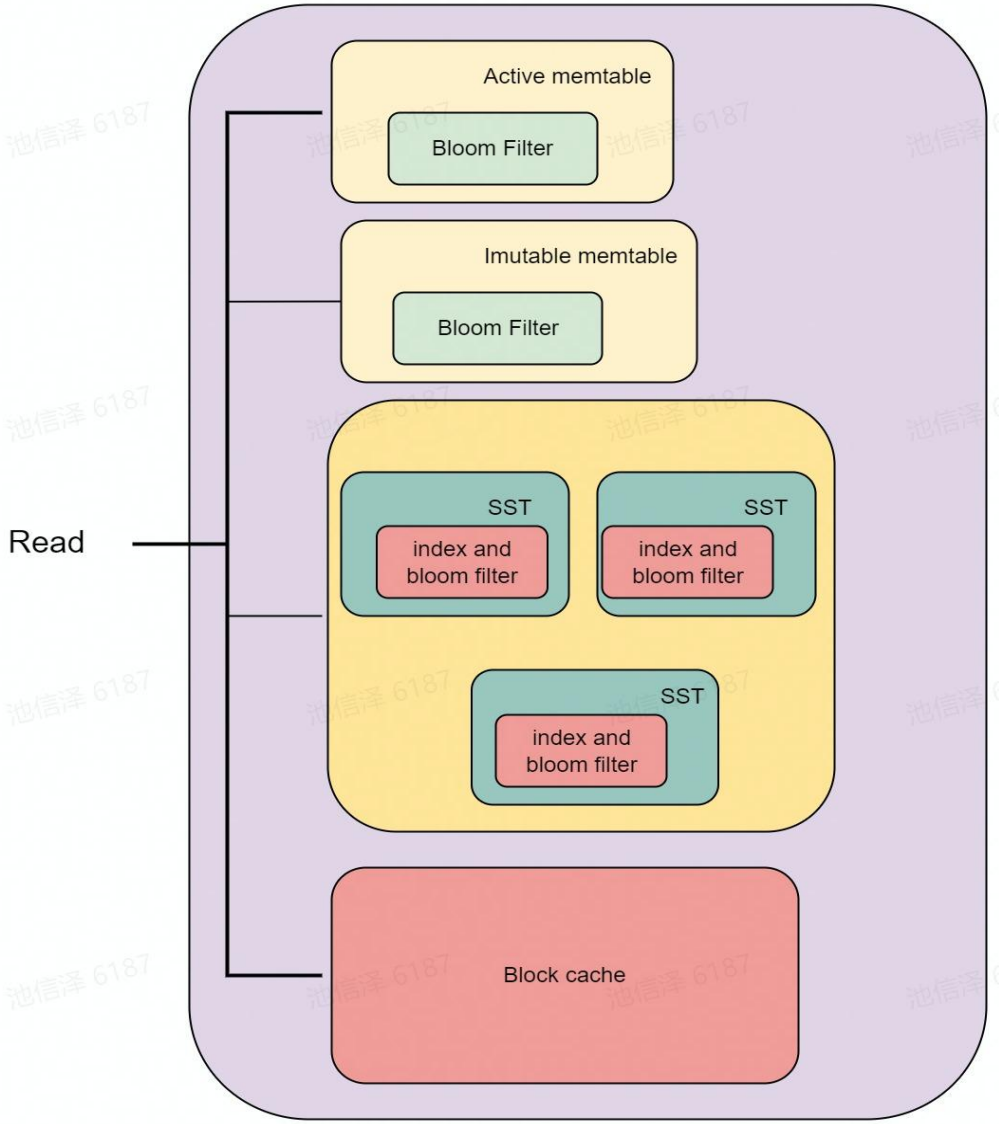
- 异步恢复
- 回填优先量恢复
- 桶自动分片
- 限制快照规格
- 利用ceph—objectstoe—tool 迁移正常的数据
- 磁盘dd 对拷，忽略错误扇区

缓存因素

RocksDB是OSD元数据持久化引擎，目前OSD对于元数据的访问采用

同步IO的方式，接口的性能直接影响到IO并发， 因此提高元数据访问速

- Memtable
  - 影响写，特别长时间大压力写
  - 需评估内存消耗
- Block cache
  - 影响读
  - 需评估内存消耗
- Compression
  - 影响写也影响读
- Bloom filter
  - 影响读，减少读放大
- cache_index_and_filter_blocks
  - 影响rocksdb内部元数据的cache策略
  - 需评估内存消耗



业务场景

常见的问题

- 大量非4K 对齐写
  - 导致所有写请求串行执行
- 顺序小IO 连续下发
  - 导致时延高
- 资源池规划粗暴
  - 导致故障半径太大

解决方案

- 一般是卷分区的时候没有4K 对齐分区
- 连续Op组合执行
- 规划合理的单个资源池规模

OpenEuler

我参与 我做主

下载体验

您可以通过以下四种方式体验openEuler操作系统

公有云　虚拟机　硬件　树莓派

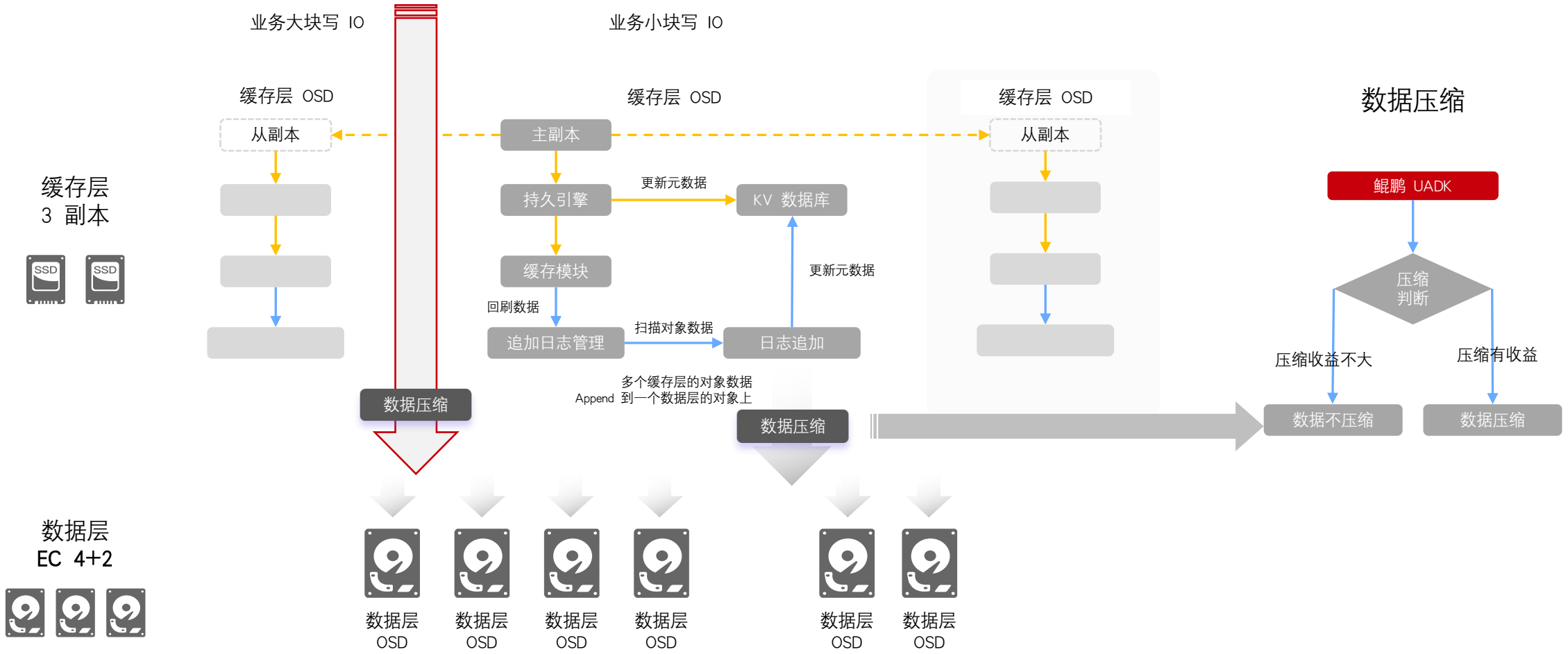
加入贡献

请根据您的参与身份，选择签署：

个人CLA　企业CLA　员工CLA

鲲鹏 UADK 使能 XSKY SDS

全场景 EC & 压缩

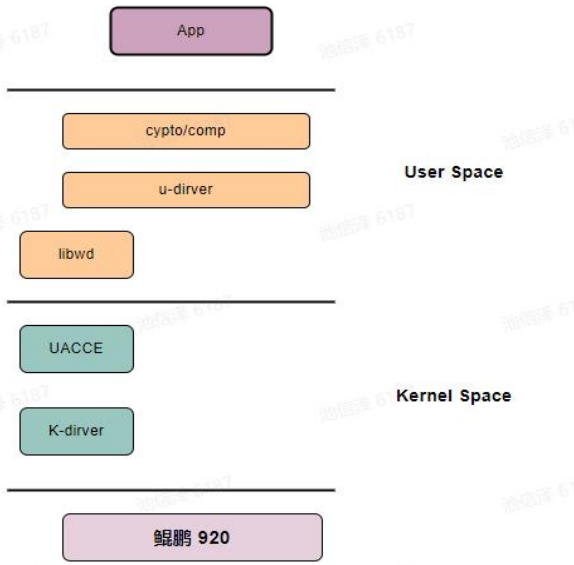

业务大块写 IO
业务小块写 IO
缓存层 OSD
缓存层 OSD
缓存层 OSD
从副本
主副本
从副本
持久引擎
更新元数据
KV 数据库
缓存模块
更新元数据
回刷数据
扫描对象数据
追加日志管理
日志追加
多个缓存层的对象数据
Append 到一个数据层的对象上
数据压缩
数据压缩
缓存层
3 副本
SSD
SSD
数据层
EC 4+2
数据层
OSD
数据层
OSD
数据层
OSD
数据层
OSD
数据层
OSD
数据层
OSD
数据压缩
鲲鹏 UADK
压缩
判断
压缩收益不大
压缩有收益
数据不压缩
数据压缩

鲲鹏UADK 的介绍

UADK 全称为 User Space Accelerator Development Kit，是一套用户态硬件加速器开发工具集，SVA 技术下用户可以高效地利用鲲鹏硬件加速器能力，为用户提供基础的库和驱动支持。

- 基于IOMMU模式限定了设备和进程的访问权限和安全边界，保证访问安全
- SVA支持设备基于用户态进程申请的VA 直接DMA 操作，不经过系统调用，内存拷贝和地址转化

Kunpeng 920 支持同步和异步模型，支持常用数据加密算法和zlib，gzip 压缩算法。



基于鲲鹏UADK 的性能优化

XSKY分布式存储基于鲲鹏计算平台的UADK 工具，大幅度提升分布式存储的性能

- XSKY 一体机使用openEuler ＋ 鲲鹏处理器
  - 4KB 随机读性能提升20%，随机写性能提升10%+，CPU 使用率平均降低了20%+
  - 512K 读性能提升30%+，512K 写性能提升20%+，CPU 使用率平均降低了30%+